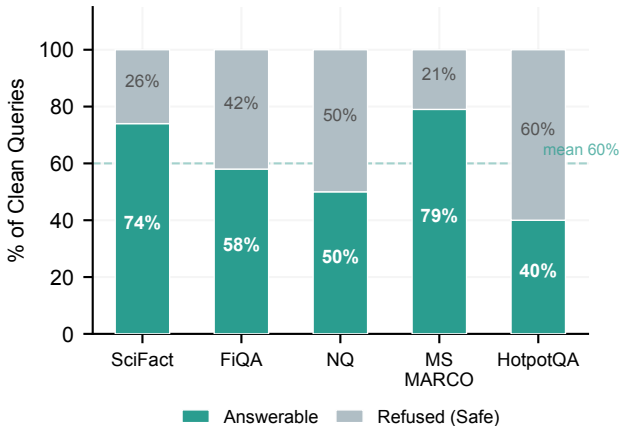


Clean Utility: Coverage-Safety Trade-off

Clean Answerability (60% mean)



Answer Correctness (LLM-Judged)

